A Socio-contextual Approach in Automated Detection of Cyberbullying

Nargess Tahmasbi
Pennsylvania State University
nvt5061@psu.edu

Elham Rastegari
University of Nebraska Omaha
erastegari@unomaha.edu

Abstract

Cyberbullying is a major cyber issue that is common among adolescents. Recent reports show that more than one out of five students in the United States is a victim of cyberbullying. Majority of cyberbullying incidents occur on public social media platforms such as Twitter. Automated cyberbullying detection methods can help prevent cyberbullying before the harm is done on the victim. In this study, we analyze a corpus of cyberbullying Tweets to construct an automated detection model. Our method emphasizes on the two claims that are supported by our results. First, despite other approaches that assume that cyberbullying instances use vulgar or profane words, we show that they do not necessarily contain negative words. Second, we highlight the importance of context and the characteristics of actors involved and their position in the network structure in detecting cyberbullying rather than only considering the textual content in our analysis.

1. Introduction

Cyberbullying has become a main threat to online social communities. It refers to a bullying conducted through an online social medium [11]. The most vulnerable target population of cyberbullying are adolescents. Reports claim that one out of five students in the United States is a victim of cyberbullying [1]. Before the introduction of online social media platforms, bullying in the physical environment used to occur at schools. The school bullies risk facing consequences from school administration.

After the introduction of online social media, bullying has become more widespread mainly because of the features of social media that facilitate spread of text and media. Unlike conventional bullying, cyberbullying does not end at schoolyards. With 73% of U.S. teens owning smartphones and 92% of them going online daily [2], it is not far from expectation that teens take the bullying to online environment after school.

Moreover, the scope of the effect of cyberbullying is much broader than that of physical bullying. The range of audience the bullies can reach in a matter of hours via online social media is far beyond than that of a schoolyard and thus the harm is more intense on the victim. Majority of research in conventional bullying attempted to identify the motivation behind bullying and looked at the problem from socio-psychological and educational perspectives.

With the increasing growth of cyberbullying incidents in recent years, scholars have attempted to study the motivational factors behind bullying in online social platforms. A majority of these studies still stem from psychology and education disciplines [3], [4]. A few computational studies have analyzed cyberbullying incidents in an attempt to automatically detect the instances. Among the computational studies of cyberbullying, most studies have assumed that cyberbullying contents usually include negative or profane words [5]–[7]. Thus they used a dictionary of bad words as a reference for comparing and identifying how similar the word vector of the cyberbullying text is similar to the vector of bad words. However, using negative words in a comment posted online is not always an indicator of cyberbullying occurrence [8]. Instead, the characteristics of the poster and their previous pattern of online behavior may serve as an indicator even though the content posted online may not contain any negative words. For example, in the collection of tweets that we have populated for this study, %4.7 of the contents are cyberbullying instances and not many instances of negative comments are present among them.

Our research objectibeve in this study is to combine the textual information with social and contextual characteristics and find the significant factors among them to propose a cyberbullying detection model. The main research question is: *what is the most significant combination of factors that lead to an accurate automatic detection of cyberbullying content?*

The socio-contextual characteristics that we investigated in our study include the characteristics of actors involved in the cyberbullying and the social network structure around the incident. We will contribute by introducing a socio-contextual approach

URI: http://hdl.handle.net/10125/50157
ISBN: 978-0-9981331-1-9

which will be proved to work better in terms of accuracy than purely textual, social, or contextual approaches. Also we demonstrate that, depending on the context, in some cyberbullying incidents, the bullying messages are not necessarily containing negative content and thus, more complex approaches are required to combine different sets of features to achieve a more accurate model.

The rest of this paper is structured as follows. First we provide a background of cyberbullying including previous studies in the area. Second, we explain our data collection and research method. We provide our results and discuss the finding in the discussion section. We conclude this paper with suggestions for future research.

2. Background

Bullying is referred to as targeted intimidation or humiliation caused by a physically or socially stronger person to make the victim powerless, threatened, or belittled [9]. To differentiate bullying from other types of aggression, Olweus has identified three criteria for bullying: intentionality, repetition, and power imbalance between the bully and the victim [10]. In the physical type of bullying, the power imbalance is an important factor distinguishing a bullying incident from other types of conflict [9]. With the advent of new computer communication tools, especially online social platforms, bullying has gained another form as known as cyberbullying. It is similar to conventional bullying in definition as it simply refers to a bullying conducted through an online social medium [11]. More specifically, Slonje and Smith defined cyberbullying as "*an aggressive, intentional act or behavior that is carried out by a group or an individual repeatedly and over time [through modern technological devices such as mobile phones or internet], against a victim who cannot easily defend him or herself*" [12].

All three criteria for defining cyberbullying suggested by Olweus [10] are applicable to the modern definition of cyberbullying [13]. Two conditions provided by the new computer mediated communication technology intensify the motivation of the bully and the negative impact of bullying on the victim. These two conditions include anonymity and public vs private dissemination of negative contents [13].

With the increasing growth of cyberbullying incidents in recent years, a significant stream of research started to make sense out of this phenomenon to provide insight on the motivation behind cyberbullying as well as to provide automated detection methods for identifying these incidents. Majority of research in this area is from sociopsychology and educational perspective and is dedicated to identifying motivations and providing mitigation solution using qualitative methods [14]–[16].

This stream of research in cyberbullying provides us insight on the cyberbullying motives and the scope of its impact on the victim and highlight the role of online social platform in facilitating cyberbullying. However, when it comes to automated detection of cyberbullying, these approaches are not suitable as their primary focus is on the mitigating phase of cyberbullying which seeks to sooth the negative impact of cyberbullying on the victim.

The abundance of data on online conversation over the internet provides us an opportunity for analysis of real life data on cyberbullying incidents. Computational studies have used quantitative methods in an attempt to automatically identify cyberbullying instances. Majority of these studies use textual features to identify the cyberbullying cases [17]–[21]. Bag-of-words is the most common method seen in the literature for identifying negative words (swear words, profane words and the like) in the corpus (e.g. [7]).

Studies with textual perspective mostly assume that cyberbullying contents include some sort of profane or in general negative words. However, identification of cyberbullying instances in most cases is more complicated than this approach. A cyberbullying content may contain non-negative words and still be cyberbullying. For example, a person might get picked up by a group of others mocking a statement he/she has made before. The mocking statements from others may not necessary have negative content, but when repeated several times by different people over time it becomes a bullying incident. Sometimes, a cyberbullying incident may start by a group of people systematically trending a hashtag on a social media platform in response to a previous incident. Identifying cyberbullying incidents, is not feasible without investigating the context of the incident.

A few studies have suggested or incorporated contextual information in their analysis [22], [23] and a few others have taken a socio-textual perspective and investigated the role of network structure in improving the detection methods [24], [25]. Understanding the social network structure can give us insight on the personality traits of users [26]. Furthermore, personality traits are reported to have correlation with cyberbullying [27]-[29]. Some of these personality traits are narcissism [27], callous-unemotional traits [28], and Dark Triad personality traits [29]. While the social network features have potential to determine some of these traits, computational studies in cyberbullying detection have mostly ignored the

personality traits and characteristics of users in predicting cyberbullying incidents.

Another gap in automated cyberbullying detection research is that not many of the studies consider the temporal dimension of the incident into their analysis. Sometimes a cyberbullying post on a social media website may not be easily identified without knowing the history of the posters' behavior and their pattern of content dissemination before the incident.

A common challenge in cyberbullying detection research is obtaining a proper dataset which contains enough cyberbullying instances for analysis. In most cases, the proportion of the cyberbullying instances is very low that leads to the problem of imbalance class distribution. Moreover, because of the lack of unanimity in definition of cyberbullying, labeling of the incidents becomes a challenging task as labelers are not confident about what constitutes a cyberbullying instance.

In this study we will address the aforementioned research gaps by proposing our data collection method and our analysis method that takes into consideration both textual and socio-contextual features in the prediction model.

3. Research Method

3.1. Data Collection

We collected our data from a stream of Tweets posted over the course of 4 days. The incident started on June 5th, 2017 after a media personality announced in a tweet that he has been blocked by a celebrity with whom he had verbal conflict recently. Soon after, the fans of the celebrity started mocking the media personality by trending a particular hashtag and mentioning him in their tweets.

We used Twitter API and Python script to collect all tweets containing the bullying hashtag that is specifically used for the purpose of cyberbullying the media personality. Total of 1790 tweets were found out of which 410 were English. We then extracted all the English speaking users involved in this cyberbullying incident. This list included all the users who tweeted at least one tweet with the cyberbullying hashtag, the users who have been mentioned in at least one of these tweets, and the users who have been retweeted at least once by other users. Then we collected all tweets from the user list that have been posted from June 3rd-6th. We waited till the end of the day of June 6th to collect the tweets to have a complete list of tweets for the last day. This step gave us 12837 English tweets which contained 607 cyberbullying tweets. 8850 were retweets from other users which contained 388 cyberbullying tweets and the remaining 3987 were original tweets (containing replies as well) containing 219 cyberbullying instances.

This approach of data collection helped us bypass the problem of data annotation and labeling which is a confusing task due to the lack of unanimity in defining what constitutes cyberbullying and subjectivity of the labeling process to the interpretation of human labelers. In this case, the cyberbullying tweets were already labeled by users by using the hashtag which was specifically designed and trended for the purpose of cyberbullying the media personality.

We consider this case as a cyberbullying case for the following reasons according to the criteria defined by Olweus [10]. First, there seems to be a power imbalance between the victim and the bullying group. While the victim has relatively high number of followers (13K at the time of data collection), the volume of tweets targeting the victim and the range of audience the cyberbullies could reach as a group were significantly higher than the range of the audience the victim could reach. Moreover, the cyberbullying group mostly comprised the fans of the celebrity who, per the victim's claim, has blocked the victim. This teens' celebrity had 96.5 Million followers at the time of data collection which is far higher than the number of followers (potentially supporters) of the victim (the media personality). This imbalance resonates the power imbalance between the bully and the victim. It is worth noting that in this case the celebrity is not the bully and the power imbalance is between the combined power of the large audience that support the celebrity and the power of the victim in defending himself.

Second, there is a repetition evident in cyberbullying of the victim. In the course of two days we have collected more than 600 cyberbullying tweets which were constantly increasing in the following hours.

Third, the last criterion of bullying is also present in this case. The act of creating a hashtag which is solely used for the purpose of mocking the victim with bullying tweets shows the intention of the group in cyberbullying the targeted individual.

Since all the three bullying criteria defined by Olweus are present in this case, we consider this case as a cyberbullying incident.

This case is also related to the cyberbullying incidents among adolescents in a way that the cyberbullying occurs in support of a teens' celebrity; thus, although the victim is not a teenager but majority of cyberbullies are in their teenage ages. Therefore, we foresee that by using this case as our dataset, we will shed lights on detection methods of cyberbullying among young generation.

3.2. Data Analysis

3.2.1. Textual features. In this study, we do not bias our perception of cyberbullying content toward contents that necessarily include negative or profane words. As mentioned before, our aim is to not make any assumption on negativity of the content as many cyberbullying cases do not include even moderate negative content. We base our analysis on general linguistic features that can be extracted from text using linguistic tools. Our selection of textual features is based on previous literature and extracted using LIWC (Linguistic Inquiry Word Count) tool.

Among the features supported by LIWC, we have selected the following to be extracted from our corpus:
(1) 'we' words. Bullying sometimes occurs in groups. Salmivalli et. al have differentiated between different roles in bullying in schools ranging from the bully, to reinforcer of the bully, to assistant of the bully [30]. Individuals in each role are usually form a group and refer to the victim as someone not belonging to their group. Similarly, cyberbullies may incorporate linguistic features to verbally reject the victim from their group. We propose that the usage of 'we' words (e.g. *we*, *our*, *us*, *let's*) as a means of expressing belongingness to group is different in cyberbullying messages and non-cyberbullying messages. On the other hand, according to the same argument, the usage of 'I' words is expected to be lower in cyberbullying messages as group cyberbullying is more about separating an individual victim from 'us' as a group, rather than 'I' in this case.
(2) 'Anger' words. Based on research studies on physical bullying, the inability to control anger is one of the characteristics associated with bullying behavior in both bullies and victims [31]–[33]. We propose that people use more 'anger' words in cyberbullying messages than that of non-cyberbullying texts. Examples of anger words include *damn*, *savage*, *hate*, and *hell*.
(3) 'Power' words. Power imbalance is identified by Olweus [10] as one of the three criteria considered for categorizing an act as bullying. Thus, it is expected that cyberbullying messages contain more 'power' words than non-cyberbullying messages. Examples of power words include *strong*, *important*, *win*, and *never*.
(4) 'Gender' words. Gender differences has been reported in cyberbullying among middle school children in which females are more victims of cyberbullying [34] meaning that more female words (e.g. *she*, *her*, *girl*) in the cyberbullying messages are expected if we are analyzing the messages among middle school children. In this case, we will investigate the usage of both 'female' and 'male' words in the two categories of tweets. However, in this particular case, it is expected that the usage of male words to be higher as the victim is a male user.
(5) 'Positive' and 'negative' words. Positive or negative tone of a message is considered as a language feature effective for cyberbullying detection [35]. Many cyberbullying detection studies claim that cyberbullying contents include negative words [5]. We investigate both negative (e.g. *sigh*, *evil*, *smh*, *fight*) and positive words (e.g. *love*, *happy*, *cutie*, *thank*) and the potential difference of tone in cyberbullying and non-cyberbullying instances.
(6) Authenticity. The main intention in cyberbullying is to make the victim feel bad and belittled. Thus, cyberbully does not necessarily believe in what he/she writes as the main point is to target the victim with a bullying message. Authenticity of a text can be measured by LIWC authentic features which is defined as '*speakers belief in the text'* [36]. Authentic sentences usually use first person pronoun and may include words such as *always*, *don't*, *think*, *true*, *better*, *though*, and *still*. We propose that cyberbullying tweets sound less authentic in general than non-cyberbullying tweets.

We present four categories of hypotheses that need to be tested. Category 1 hypotheses pertains to the association between textual features and cyberbullying nature of tweets. We define this hypothesis as:
H1. Textual characteristics of cyberbullying tweets is different than that of non-cyberbullying tweets.
We have defined sub-hypotheses that help us test the main hypothesis with objective measures. The hypotheses included in category 1 are as follow:

H1-a. Cyberbullying tweets use more 'we' words on average than non-cyberbullying tweets.

H1-b. Cyberbullying tweets use less 'I' words on average than non-cyberbullying tweets.

H1-c. Cyberbullying tweets use more anger words on average than non-cyberbullying tweets.

H1-d. Cyberbullying tweets use more power words on average than non-cyberbullying tweets.

H1-e. Average usage of gender words in cyberbullying tweets is different than that of non-cyberbullying tweets.

H1-f. Cyberbullying tweets use less positive words on average than non-cyberbullying tweets.

H1-g. Cyberbullying tweets sound less authentic on average than non-cyberbullying tweets.

3.2.2. Network features. We propose that users' network structure is relevant to the users' spread of bullying content on Twitter. Studies have confirmed that network structure can be used to identify personality traits [26], [37]; and personality traits, on the other hand, have correlations with the user's behavior on social networks and specifically the act of

committing cyberbullying [27]–[29]. More specifically, for instance, degree centrality is reported to have high correlation with extraversion [37]. In another study, betweenness centrality is proved to be associated with conscientiousness, extraversion, and neuroticism, while closeness and degree centralities are correlated with age in addition to all of the above [26]. Among the centrality measures, degree centrality is the simplest one. It refers to the number of other elements in the network that are connected to the current element [38]. In a directed network, where the direction of a tie matters, one can differentiate between the number of incoming and outgoing ties and call them in-degree and out-degree respectively. In Twitter social network, degree centrality can be measured in different ways. The number of followers a user has or the number of retweets or mentions a user receives can be indicators of in-degree centrality. And vice versa, the number of users a person follows or the number of retweets or replies the user makes to other users can be indicators of out-degree centrality. Betweenness centrality is a measure that determines the power of an individual in a network in terms of how often he/she can interrupt the flow of information or how often the person acts as a mediator of communication between any other two individuals in the network. Closeness centrality is determining how often the user can bypass the mediators to reach to the other users in a shorter number of steps. In the Twitter space, this can be translated into how many retweets or mentions in a row (on average) can take the user to another user in the network. In our analysis, we measure all three centralities mentioned above from the retweet activity viewpoint. We calculate the centrality measures of all users based on their retweet network during two days before the cyberbullying hashtag started becoming trending. We did not include the centrality measures affected by the users' activity after the incident started as we are interested to investigate the current status of the users in the network and its correlation with their future behavior and its prediction power in identifying the cyberbullying posting.

Category 2 hypotheses are developed to identify the association between social network features and cyberbullying nature of tweets. The main hypothesis for this category is:

H2. Average network measures of posters is different in cyberbullying and non-cyberbullying tweets.

Sub-hypotheses included in this categories are as follow:

H2-a. Average degree centrality of posters is different in cyberbullying and non-cyberbullying tweets.

H2-b. Average closeness centrality of posters is different in cyberbullying and non-cyberbullying tweets.

H2-c. Average betweenness centrality of posters is different in cyberbullying and non-cyberbullying tweets.

3.2.3. Meta-features. Pictures and video clips bullying are reported to have more negative impact on the victim [3]. We have checked for the presence of any type of media (picture/video clip) in the tweets to identify the potential role of media usage in identifying cyberbullying contents. Moreover, we have intention to investigate if the cyberbullying contents are more conversational in nature than non-cyberbullying contents and if this measure can have prediction power in identifying cyberbullying tweets. Thus, we extract the number of users that have been mentioned or replied to in the tweet content. This measure can serve as an indicator of how many people are engaged in the conversation carried over by a tweet post. Other tweet meta-features that are included in our analysis are related to the tweet's popularity. This feature is measured by the number of favorites and number of retweets a tweet receives. We intend to investigate if there is any difference between the average popularity of cyberbullying and non-cyberbullying contents and if it can be a predicting measure for identifying cyberbullying cases.

Category 3 hypothesis is proposed to test for the association between tweet metadata and cyberbullying nature of tweets. Our hypothesis is as follows:

H3. Tweet metadata features are different in cyberbullying than non-cyberbullying tweets.

Sub-hypotheses included in this category are as follow:

H3-a. The average of tweet media count is different in cyberbullying than non-cyberbullying tweets.

H3-b. The average of tweet mention count is different in cyberbullying than non-cyberbullying tweets.

H3-c. The average of tweet retweet count is different in cyberbullying than non-cyberbullying tweets.

H3-d. The average of tweet favorite count is different in cyberbullying than non-cyberbullying tweets.

In addition to tweets meta-features, we have also considered users' meta-features in our analysis. These features include user's number of friends and followers, current total number of tweets posted by user, and current total number of tweets liked by user. We calculated the ratio of the first two measures to achieve an index for the user's level of power. The more the number of user's followers compared to friends, the more indicative of the user's power in the network. This ratio can also be considered as the user's

centrality in the following/followers network. The last two features are indicative of user's activity in the social network and openness/friendliness of the user toward others.

Category 4 introduces a hypothesis regarding the association between user metadata and cyberbullying nature of tweets and is defined as follows:

H4. User metadata are different in cyberbullying than non-cyberbullying tweets.

The sub-hypotheses to test the category 1 hypothesis are as follow:

H4-a. The average users' ratio of followers to friends is different in cyberbullying than non-cyberbullying tweets.

H4-b. The average users' total number of tweets is different in cyberbullying than non-cyberbullying tweets.

H4-c. The average total number of tweets liked by the user is different in cyberbullying than non-cyberbullying tweets.

3.2.3. Imbalance class distribution. As mentioned in the data collection section, the percentage of cyberbullying instances to non-cyberbullying ones in our data set is less than %5. Out of 3987 original tweets only 219 were cyberbullying instances. This leads to the problem of imbalance class distribution which may negatively affect the accuracy of prediction models. There are some resolutions for this issue mentioned in the literature. One of them is Synthetic Minority Over-sampling Technique (SMOTE) which is appropriate when there is only a few instances of the positive cases [39]. We have used SMOTE in the preprocessing step to account for the imbalance class distribution to prepare the data for classification techniques explained in the next section.

3.2.4. Classification methods. Before applying classification methods on the data, we investigated the most influential features to include in the classification process. Information gain is a frequently used feature selection method for text classification. But it can be employed for selection of different types of features as well. It works by measuring the decrease in entropy in the presence and absence of the feature [40]. We used information gain evaluation on the feature set combined with ranker method to extract and rank the most influential features which may have predication power in classifying the tweets into cyberbullying and non-cyberbullying cases.

We performed different methods of classification including Naïve Bayes, SVM, Random Forest, logistic, JRip, and J48 using a 10-fold cross validation method and compared the accuracy of each model. Then we picked the most well performing method and repeated the classification separately on each set of features: textual, network, tweet meta-features, and user meta-features.

4. Results

4.1. Inferential Statistics

We performed an independent sample t-test on all three sets of features (textual, network, and metadata) to compare the means between two groups of cyberbullying and non-cyberbullying tweets. We found that the number of words associated with *we*, *anger*, *power*, and *male* are significantly greater in bullying messages compared to non-bullying messages, while the number of words associated with *personal pronouns*, *I*, *female*, *authenticity*, *emotional tone*, and *positive emotion* are significantly less in bullying messages.

Our results confirm that cyberbullying messages have less emotional tone and positive emotion compared to non-cyberbullying messages (α=.05). However, we did not see significant difference between cyberbullying and non-cyberbullying messages regarding the negative content, meaning that bullying messages might have less positive content, but not necessarily more negative content. While not all sub-hypotheses in category 1 are supported, still a few of them are supported which confirm the support for H1.

Among the network features, closeness and betweenness centralities are reported significant (α=.05) with both measures lower for cyberbullying tweets than non-cyberbullying tweets. This confirms that H2 hypothesis is supported.

From the tweet meta-features, H3-a and H3-b were supported (α=.05). The mentions count in cyberbullying tweets was significantly less than that of non-cyberbullying tweets while the media count was significantly higher. The supported sub-hypotheses in category 3 confirm that H3 is supported.

All user meta-features were significant (α=.05). Followers/Friends ratio of the poster was significantly higher in cyberbullying tweets than non-cyberbullying tweets. In addition, current total number of tweets posted by user and the number of tweets that the user favorited were lower for cyberbullying tweets than that of non-cyberbullying tweets. This result confirms the H4 is a valid hypothesis.

4.2. Classification

Before performing the classification, we have ranked all the features according to their information

gain to obtain a set of features that are potentially significant in predicting cyberbullying instances. Table 1 shows the top 14 features used in the classification methods ranked based on their information gain, along with a brief description of each feature. The user meta-features made it to the top of the list along with most of the network features while tweet meta-features and textual features are ranked lower.

Table 1. Top features selected based on information gain

Feature	Description	Feature category
User's favorites count	Number of tweets the user has favorited (liked)	User meta-feature
User's Tweet count	Number of Tweets the user has posted	User meta-feature
Followers/friends ratio	The ratio of the number of followers to the number of people the user follows	User meta-feature
Closeness centrality	The degree of closeness of the user to other users in terms of their ability of disseminating tweets to the target audience	Network
Betweenness centrality	The degree of the being able to interrupt the flow of information and act as an information broker in the network	Network
Retweet count	Number of times the tweet has been retweeted	Tweet meta-feature
Out-Degree centrality	Number of retweets the user has made	Network
Tweet favorite count	Number of times the tweet has been liked	Tweet meta-feature
Power words	Number of power words used in the tweet (e.g. superior, bully)	Textual
'I' words	Usage of 1st person singular words (e.g. I, me, mine)	Textual
Mentions count	Number of users mentioned (replied to, mentioned, or retweeted) in the tweet	Tweet meta-feature
Female	Usage of female references (e.g. girl, her, mom)	Textual
'We' words	Number of 1st person plural words (e.g. we, us, our)	Textual
Authentic words	Speaker's belief in the text (e.g. always, don't, think, true, better)	Textual

We performed classification methods on the 14 features shown in table 1. Among the classification methods that we used, J48, and JRip, and Random Forest had the best overall performance while logistic methods, Naïve Bayes, and SVM had the worst performance. Among the top three best performing classifiers, J48 has slightly better recall for cyberbullying cases, while Random Forest has better precision for cyberbullying cases and higher accuracy overall.

Table 2. Comparison of classifiers' performance

	Accuracy	Precision	Recall	ROC Area	Precision for Cyberbullying	Recall for Cyberbullying
J48	93.91	93.6	93.9	86.2	73.2	63.9
JRip	93.78	93.4	93.8	78.4	74.8	59.4
Random Forest	95.38	95.2	95.4	95	89.8	62
Logistic	89.61	85.4	89.7	80.8	44.8	03.1
AdaBoost	90.74	89	90.7	87.4	67.5	18.6
SVM	89.75	80.6	89.8	84.9	0	0
Naïve Bayes	43.09	89.3	43.1	76.4	14.5	92.9

Table 2 summarizes the evaluation of different classification methods according to their accuracy, precision, and recall.

Among the three best performing methods, Random Forest is selected as the best method due to its higher accuracy and ROC area, as well as overall precision and recall. Thus, we select Random Forest and apply this method to each category of features explained in the previous section.

Table 3 shows the result of applying Random Forest on these categories. As the results show, using all categories as classifier features in the classification method increases the accuracy of the classification as well as the precision and recall especially for cyberbullying instances.

Table 3. Comparison of Random Forest classifier's performance on each feature category

	Accuracy	Precision	Recall	ROC Area	Precision for Cyberbullying	Recall for Cyberbullying
Textual features	92.63	90.4	91.6	78.8	69.7	32.5
Network features	90.77	89.1	90.8	79.1	69.1	17.9
User meta-features	92.97	92.2	93	88.5	74.5	47.6
Tweet meta-features	94.57	92.7	94.6	65.3	49.1	12.3
All features	95.38	95.2	95.4	95	89.8	62

5. Discussion

We approached the problem of automated cyberbullying detection of cyberbullying starting with an inferential analysis. With this analysis, we intended to show that there are differences between cyberbullying and non-cyberbullying tweets based on their usage of three categories of features. Our results supported all hypotheses proposed for textual features, network features and meta-features except for the last two sub-hypotheses in tweet meta-feature category. The first two sub-hypotheses related to tweet meta-features were related to the conversational nature of tweets. The results from the inferential statistics show that cyberbullying tweets are more conversational than non-cyberbullying tweets.

Moreover, cyberbullying tweets use more multimedia contents (image/video) than non-cyberbullying tweets which is in line with the common practice of cyberbullies especially in photo-sharing social platforms (e.g. Instagram) in which the bully posts a distorted image of the victim with a bullying message captioned on it. But at the same time, our results found no evidence supporting the assumption that cyberbullying tweets are more or less popular than non-cyberbullying tweets. This might be more related to the fact that the cyberbullying tweets in our case were spread in a short period of time (two days) that the cyberbullying tweets did not yet get a chance to get favorited or retweeted by others.

Network features were also among the ones that were significantly different in cyberbullying and non-cyberbullying instances. The results show that all three network features are significantly higher in cyberbullies. As suggested by Staiano et al., network centrality of an actor can be associated with the actor's personality trait especially the social power of the actor [26]. Based on our results, we can infer that users who cyberbullied feel more powerful on average than those who did not cyberbully.

The results of t-test show that all user meta-features were significant. These features are categorized into three classes of user's popularity/power, user's activity, and user's friendliness/openness. Hypothesis H4-a, which is relate to user's popularity/power is supported showing that the users who cyberbullied are more popular or feel more powerful on average than the users who did not cyberbully. H4-b indicates that users who cyberbullied are less active in general than the users who did not cyberbully. In addition, the support of H4-c indicates that the users who cyberbullied are less open to like other users' tweets.

This inferential analysis gives us an insight on different nature of cyberbullying and non-cyberbullying tweets. We took the step further to investigate the influential factors that have prediction power to classify tweets into cyberbullying and non-cyberbullying categories.

We obtained the most influential features using an information gain based feature selection method. Results show that user meta-features are the most influential features that have discriminatory power to predict the cyberbullying nature of a tweet, with network features and tweet meta-features in the second place while the textual features were at the bottom of the list. This indicates that not only socio-contextual features are important in automated detection of cyberbullying but they are even more important than textual features in this case.

However, we believe that some of these features are not independent from the context. For example, while studies have claimed that more female words in the cyberbullying messages are expected among middle school children, in our case study, the number of words associated with 'female' is significantly lower in cyberbullying messages compared to non-cyberbullying messages. This observation is due to the data set collected using a specific hashtag that targets a male victim.

While this study targets a specific case of cyberbullying on Twitter triggered from a conflict between a media personality and a teens' celebrity, the outcome is informative for future cyberbullying studies. The contribution of our paper is two-fold. First, as illustrated in table 3, we have shown that considering all three categories of features in the classification model significantly increases the accuracy, precision, and recall of the classification model. To the best of our knowledge, no study has incorporated all the features including network features

in the automated detection of cyberbullying. We have filled this gap by emphasizing the importance of socio-contextual features in cyberbullying detection.

Second, we broke the assumption seen in previous studies that cyberbullying texts are of highly negative and profane nature. As shown in the t-test results, there was no evidence showing the difference between content of cyberbullying and non-cyberbullying tweets in terms of negative words usage. They do however differ in terms of positive words usage. While the positive words used in cyberbullying tweets were significantly lower than that of non-cyberbullying tweets, this is not necessary inferring that cyberbullying tweets contain more negative or profane words.

This study has some limitations. First of all, we studied a specific case of cyberbullying which pertains to a celebrity case and therefore the results of our study may not be fully applicable to other cases of cyberbullying in general. However, independent of the context, consideration of all feature categories in the analysis seems to improve the accuracy of automated cyberbullying detection model. In future, we will apply the current methodology to other context to validate and extend the methodology.

Cyberbullying comes in several forms and is conducted through various online social media platforms. Future studies can take a cross-context and cross-platform approach to automated cyberbullying problem to achieve a more general solution independent of the context and medium.

Another limitation in our study is that in our data collection process, we ignored other types of media such as image and video and only extracted the textual part of the tweet. Image and video features can be equally powerful as textual and contextual features in identifying cyberbullying cases. Future studies can use image processing techniques to automatically extract features from multimedia content and incorporate them in their classification method to improve the accuracy of the model.

Another perspective to look at the cyberbullying detection problem is an actor-based detection approach in which cyberbullies are identified instead of cyberbullying contents. According to Salmivalli et al., different roles may be engaged in cyberbullying, including the bully, reinforcer of the bully and assistant of the bully [30]. These roles can be identified by screening the profiles and previous activities of the users in the social media. Our future research plan is to perform a longitudinal analysis that gives us more information about the pattern of users' previous activities and their position in the network. These features have the potential to identify the future cyberbullies based on the information on the history of current cyberbullies.

6. Conclusion

In this study, we took a socio-contextual approach to develop a model to automatically detect cyberbullying cases. According to our findings we contributed to research by concluding that cyberbullying instances do not necessarily contain profane and negative words and other than textual features, characteristics of users and their previous position in the network play an important role in differentiating between cyberbullying and non-cyberbullying instances.

7. References

[1] D. Lessne and C. Yanez, "Student Reports of Bullying: Results from the 2015 School Crime Supplement to the National Crime Victimization Survey. Web Tables. NCES 2017-015.," *Natl. Cent. Educ. Stat.*, 2016.

[2] A. Lenhart, "Teen, Social Media and Technology Overview 2015," Apr. 2015.

[3] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: Its nature and impact in secondary school pupils," *J. Child Psychol. Psychiatry*, vol. 49, no. 4, pp. 376–385, 2008.

[4] V. H. Wright, J. J. Burnham, T. I. Christopher, and N. O. Heather, "Cyberbullying: Using virtual scenarios to educate and raise awareness," *J. Comput. Teach. Educ.*, vol. 26, no. 1, pp. 35–42, 2009.

[5] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of Textual Cyberbullying.," *Soc. Mob. Web*, vol. 11, no. 02, 2011.

[6] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, Q. Lv, and S. Mishra, "Towards Understanding Cyberbullying Behavior in a Semi-Anonymous Social Network," *ArXiv14043839 Phys.*, Apr. 2014.

[7] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying: query terms and techniques," in *Proceedings of the 5th annual acm web science conference*, 2013, pp. 195–204.

[8] F. Mishna, C. Cook, T. Gadalla, J. Daciuk, and S. Solomon, "Cyber bullying behaviors among middle and high school students.," *Am. J. Orthopsychiatry*, vol. 80, no. 3, pp. 362–374, 2010.

[9] J. Juvonen and S. Graham, "Bullying in Schools: The Power of Bullies and the Plight of Victims," *Annu. Rev. Psychol.*, vol. 65, no. 1, pp. 159–185, Jan. 2014.

[10] D. Olweus, "Bullying at School," in *Aggressive Behavior*, L. R. Huesmann, Ed. Springer US, 1994, pp. 97–130.

[11] P. Gradinger, D. Strohmeier, and C. Spiel, "Definition and Measurement of Cyberbullying," *Cyberpsychology J. Psychosoc. Res. Cyberspace*, vol. 4, no. 2, Dec. 2010.

[12] R. Slonje and P. K. Smith, "Cyberbullying: Another main type of bullying?," *Scand. J. Psychol.*, vol. 49, no. 2, pp. 147–154, Apr. 2008.

[13] E. Menesini *et al.*, "Cyberbullying Definition Among Adolescents: A Comparison Across Six European Countries," *Cyberpsychology Behav. Soc. Netw.*, vol. 15, no. 9, pp. 455–463, Sep. 2012.
[14] A. Nocentini, J. Calmaestra, A. Schultze-Krumbholz, H. Scheithauer, R. Ortega, and E. Menesini, "Cyberbullying: Labels, Behaviours and Definition in Three European Countries," *J. Psychol. Couns. Sch.*, vol. 20, no. 2, pp. 129–142, Dec. 2010.
[15] S. Hinduja and J. W. Patchin, "Cyberbullying: An Exploratory Analysis of Factors Related to Offending and Victimization," *Deviant Behav.*, vol. 29, no. 2, pp. 129–156, Jan. 2008.
[16] V. Šléglová and A. Cerna, "Cyberbullying in Adolescent Victims: Perception and Coping," *Cyberpsychology J. Psychosoc. Res. Cyberspace*, vol. 5, no. 2, Dec. 2011.
[17] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, pp. 1–30, Sep. 2012.
[18] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proceedings of the 17th International Conference on Distributed Computing and Networking*, 2016, p. 43.
[19] C. Van Hee *et al.*, "Automatic detection and prevention of cyberbullying," in *International Conference on Human and Social Analytics (HUSO 2015)*, 2015, pp. 13–18.
[20] L. P. Del Bosque and S. E. Garza, "Aggressive text detection for cyberbullying," in *Mexican International Conference on Artificial Intelligence*, 2014, pp. 221–232.
[21] B. S. Nandhini and J. I. Sheeba, "Cyberbullying Detection and Classification Using Information Retrieval Algorithm," 2015, pp. 1–5.
[22] P. Galán-GarcÍa, J. G. De La Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," *Log. J. IGPL*, p. jzv048, 2015.
[23] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in *European Conference on Information Retrieval*, 2013, pp. 693–696.
[24] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Comput. Hum. Behav.*, vol. 63, pp. 433–443, 2016.
[25] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, "Identification and characterization of cyberbullying dynamics in an online social network," 2015, pp. 280–285.
[26] J. Staiano, F. Pianesi, B. Lepri, N. Sebe, N. Aharony, and A. Pentland, "Friends don't lie: inferring personality traits from social network structure," 2012, p. 321.
[27] F. Eksi, "Examination of Narcissistic Personality Traits' Predicting Level of Internet Addiction and Cyber Bullying through Path Analysis.," *Educ. Sci. Theory Pract.*, vol. 12, no. 3, pp. 1694–1706, 2012.
[28] K. A. Fanti, A. G. Demetriou, and V. V. Hawa, "A longitudinal study of cyberbullying: Examining riskand protective factors," *Eur. J. Dev. Psychol.*, vol. 9, no. 2, pp. 168–181, Mar. 2012.
[29] S. Pabian, C. J. S. De Backer, and H. Vandebosch, "Dark Triad personality traits and adolescent cyber-aggression," *Personal. Individ. Differ.*, vol. 75, pp. 41–46, Mar. 2015.
[30] C. Salmivalli, K. Lagerspetz, K. Björkqvist, K. Österman, and A. Kaukiainen, "Bullying as a group process: Participant roles and their relations to social status within the group," *Aggress. Behav.*, vol. 22, no. 1, pp. 1–15, 1996.
[31] A. M. Candelaria, A. L. Fedewa, and S. Ahn, "The effects of anger management on children?s social and emotional outcomes: A meta-analysis," *Sch. Psychol. Int.*, vol. 33, no. 6, pp. 596–614, Dec. 2012.
[32] K. M. Champion, "Victimization, anger, and gender: Low anger and passive responses work," *Am. J. Orthopsychiatry*, vol. 79, no. 1, pp. 71–82, 2009.
[33] P. J. Lovegrove, K. L. Henry, and M. D. Slater, "Examination of the predictors of latent class typologies of bullying involvement among middle school students," *J. Sch. Violence*, vol. 11, no. 1, pp. 75–93, 2012.
[34] M. Pujazon-Zazik and M. J. Park, "To Tweet, or Not to Tweet: Gender Differences and Potential Positive and Negative Health Outcomes of Adolescents' Social Internet Use," *Am. J. Mens Health*, vol. 4, no. 1, pp. 77–85, Mar. 2010.
[35] H. Lieberman, K. Dinakar, and B. Jones, "Let's gang up on cyberbullying," *Computer*, vol. 44, no. 9, pp. 93–96, 2011.
[36] M. Dalvean, "Changes in the style and content of Australian election campaign speeches from 1901 to 2016: A computational linguistic analysis," *ICAME J.*, vol. 41, no. 1, Jan. 2017.
[37] S. Wehrli and others, "Personality on social network sites: An application of the five factor model," *Zurich ETH Sociol. Work. Pap. No 7*, 2008.
[38] M. E. Newman, "The mathematics of networks," *New Palgrave Encycl. Econ.*, vol. 2, no. 2008, pp. 1–12, 2008.
[39] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM Sigkdd Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, 2004.
[40] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, no. Mar, pp. 1289–1305, 2003.